DOCUMENT RESUME

ED 360 839 FL 021 416

AUTHOR Gill, Martin
TITLE The Significance of "Significance."
REPORT NO ISSN-0959-2253
PUB DATE 93
NOTE 20p.; For serial publication in which this paper appears, see FL 021 410.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Journal Articles (080)
JOURNAL CIT Edinburgh Working Papers in Applied Linguistics; v4 p63-80 1993

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Applied Linguistics; Case Studies; Foreign Countries; *Linguistic Theory; *Research Methodology; *Statistical Analysis; *Statistical Significance; *Testing

ABSTRACT

This paper examines some of the implications of testing for statistical significance. After considering methodological issues raised by two examples from the literature, the paper proceeds to look in detail at a variety of misunderstandings attached to the reporting of "significant" results. It is concluded that significance testing is of limited utility and highly misleading. Implications of abandoning the significance test are considered. (JL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 360 839

The Significance of "Significance"

Martin Gill (DAL)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Brian Parkinson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

BEST COPY AVAILABLE

FL 021412

THE SIGNIFICANCE OF "SIGNIFICANCE"

Martin Gill (DAL)

Abstract

Testing for statistical significance is an integral part of the methodology of research in applied linguistics, yet its implications are easily neglected. This paper examines some of them. After considering methodological issues raised by two examples from the literature, it proceeds to look in detail at a variety of misunderstandings attached to the reporting of "significant" results. Its conclusion is that significance testing is, at best, of limited utility, but, as commonly used, highly misleading. A final section considers the implications of abandoning the significance test.

1. Scientific versus statistical reasoning: two case studies

1.1 Introduction

Applied linguists, like other researchers in the human sciences, typically look to experiments to provide the hard data necessary to corroborate their hypotheses. Despite regular calls for greater use of qualitative procedures, drawn chiefly from ethnography, the field's dominant methodological paradigm, or the one to which it aspires, continues to be that of experimental science. As in physics, results obtained by experiment, properly insulated against sources of error, are taken (ideally) to permit valid inference to principles operating in the world; and the designers of such experiments rarely hesitate to claim that a decision to reject the null hypothesis (H_0), triggered by a result at the .05 level of significance, substantively strengthens not only the particular alternative hypothesis (H_1) they favour, but the theory from which it was derived. They may also claim to have inaugurated a promising research programme, to be carried forward as a matter of urgency with further detailed study, replication with larger samples, etc.

It is the aim of this paper to examine the soundness of such convictions; in particular to note points at which an analogy with physical science may misrepresent the experimental activity of applied linguistics, and so introduce the potential for distortion into the design and interpretation of applied linguistics research. This will involve taking a close look at the nature of the test for statistical significance and various false inferences that may be drawn from it. The paper also attempts to show how a more general tendency to regard statistical procedures as scientific instruments for uncovering independently existing empirical phenomena obscures the more basic question of what research objects are, and the role of a methodology in defining them. It is argued that, to the extent that a methodology is constitutive of the kinds of knowledge it makes possible, the notion of 'independently existing phenomena' cannot

EDINBURGH WORKING PAPERS IN APPLIED LINGUISTICS, number 4 (1993) ISSN 0959-2253

be sustained. We turn first to two examples which serve to illustrate some of these issues.

1.2 Supporting a theory (Hafiz and Tudor 1989)[1]

Hafiz and Tudor report the successful results of a 12-week reading programme 'inspired by Krashen's Input Hypothesis' (Hafiz and Tudor 1989:4) which they conducted using simplified readers with a class of young Pakistani ESL learners: on post-testing, the experimental group achieved statistically significant improvements on their pre-test scores on all parts of a battery of reading and writing tests, while, for the most part, the two control groups did not. This outcome may not strike the outsider as especially surprising, given that the participants in the programme had devoted on average 42 extra hours to English, not counting the reading they did at home (ibid.:7), whereas the control groups had merely followed their usual classes. Yet for Hafiz and Tudor it

> lend[s] support to Krashen's Input Hypothesis, indicating that extensive L2 input in a tension-free environment can contribute significantly to the enhancement of learners' language skills, both receptive and productive.
>
> (Hafiz and Tudor, op. cit.:10)

Remarks of this kind are common enough in the literature. They create an impression of scientific progress, of theoretical understanding fortified by the accumulation of well-attested empirical results. But they ignore crucial differences between the logic of scientific reasoning and that of the probabilistic reasoning (as institutionalized in the t-test and its more elaborate extensions) from which the data they are concerned with are derived. These differences, therefore, deserve to be clearly stated.

What is wrong with the conclusion Hafiz and Tudor draw from their experiment has nothing to do with the plausibility of the suggestion that tension-free reading can help learners to become more proficient; nor is it just a consequence of the shakiness of their experimental design, although of course that is relevant to judging the validity of their research as a whole (cf. section 3.4 below). It is the invalid inference they make (or imply by their choice of words) from specific observation to general theory. It is a mistake that is easily overlooked in a procedure which applies statistical significance tests for the purpose of adjudicating between hypotheses, for this practice obscures the fact that the decision so determined is logically independent of any inference to the truth or otherwise of the theory from which the hypotheses are deduced, or of any attempt to attach a degree of confirmation or even remote probability to it. The results Hafiz and Tudor obtained may well be consistent with Krashen's Input Hypothesis, but they lend no more support to it than to any other plausible (or, for that matter, any wholly absurd) theory with which they might also happen to be consistent.

It is easy to see that this experimental programme was designed to illustrate a conclusion to which the researchers were already committed. Probably this increased the practical value of the programme for the group that took part in it, but it negates any scientific claim to have tested a theory, or corroborated general principles of language learning. If this escaped the researchers, it was perhaps because the significance test procedure in itself appeared to be sufficient guarantee of a scientific outcome. Notwithstanding the absence of the 'testable and falsifiable universal laws and initial conditions' that Popper (1979:193) sets as a precondition of scientific explanation, an experimental result achieving statistical significance is likely to seem (especially to

anyone already persuaded) a persuasive indication of the existence of a fact of "genuine", intrinsic significance (we may speculate which sense of this word is intended in the quotation above). This can occur, as here, regardless of doubts about the experimental design and sampling procedure, regardless of the looseness of the operational definitions chosen, and regardless of how short of specific predictive power the theory in question may be.

It would seem that any analogy between this procedure and physics must be mistaken; the credibility of physical theories is not increased by decisions of the sort determined by significance testing. But nor is it increased by observing a non-zero difference in the predicted direction between experimental and control conditions, at least not if the observation in question is entirely consistent with everyday expectations (for example, that learners learn better when they are relaxed, interested, etc.). What is required of a theory is a capacity to make novel predictions which can be subjected to exact empirical scrutiny. A theory can be said to be strengthened, at least our belief in it can be said to be more adequately justified, the longer it survives the closest scrutiny we can give it. This presupposes a theory with some interesting empirical content, capable of refutation: to the extent that Krashen's is not such a theory (as Gregg (1984) and McLaughlin (1987) argue it is not), Hafiz and Tudor could not have hoped to lend it a crumb of support, whatever their method.

Given the inherently inexact nature of their subject, very few theories in the behavioural sciences are likely to measure up to these "scientific" criteria, for example by successfully predicting the size (not just the direction) of a difference. The main justification for using the significance test is to fill this absence of precision by supplying a way of deciding when an observed value is unlikely to have occurred by chance. However, the dangers exemplified here are, first, that "achieving significance" may seem to play a role in the logic of theory-testing and the rhetoric of research-paper writing that is equivalent to that of overcoming the much more demanding observational hurdles usual in physical science; and, second, that a theory without substance will be dignified, and its position consolidated, by the published announcement of "confirmation". The logical problem of confirmation is discussed further in section 3.1, and its relation to theory-testing in the physical sciences is developed in 3.5.

1.3 Exploring a concept (Ferguson and Maclean 1991)[2]

Experiments in the behavioural and human sciences, including applied linguistics, often proceed in an "exploratory" fashion, without a theoretically motivated design, but trusting to statistical techniques to reveal what phenomena are of interest. Given the complexity of many of these techniques, it becomes tempting to view this as purely an instrumental matter, the atheoretical application of sophisticated tools to get at the underlying constituents of reality (the "facts") on the basis of which a subsequent theory will be constructed. Here too there may persist some imagined parallel with what physicists do. Not only is this image misleading, however (for physical no less than for statistical sciences), it also leaves the experimenter unguided as to which phenomena might be genuine, and which simply artefacts of the chosen method. As John Dewey observed:

> A quantitative statement with no theory to determine what is being measured would justify calling the "measuring" of all cracks in the plaster of my wall "science" if it were done with elaborate statistical technique.
>
> (Dewey 1949; cited in Johanningmeier 1980:54)

In a recent study, Ferguson and Maclean (1991) seek to analyse the properties of subjective judgements of (medical) text difficulty, and, in particular, by Principal Components Analysis, 'to get below the surface of things' (op. cit.:123) to establish the (true) dimensionality underlying the seven explicitly formulated categories of difficulty used by their team of judges in assessing texts. It emerges from the computation that there are just 'two significant dimensions' of difficulty (ibid.:122), which therefore, it seems, are to be regarded as the real, unconscious causes of the judges' conscious behaviour: 'Perhaps, then, the judges were in fact operating with two dimensions though they may have believed they were independently assessing seven' (ibid.). Applying the statistical procedure has not merely revealed broad patterns of co-occurrence in the data, but got at hidden facts which are in some sense intrinsically more explanatory than those on the surface (this may explain why the writers do not report the views of the judges themselves about the judging task). Moreover, the analysis assumes that the sense in which these facts are more explanatory is cognitive, equating their hiddenness in the data with the hiddenness of mental activity in the heads of the judges.

We might recall J.S. Mill's warning about the dangers of reification:

> The tendency has always been strong to believe that whatever received a name must be an entity or being, having an independent existence of its own. And if no real entity answering to the name could be found, men did not for that reason suppose that none existed, but imagined that it was something peculiarly abstruse and mysterious.
>
> (J.S. Mill cited in Gould 1981:320)

What is striking in this context is that the entities in question had no name before this particular study found them, but that, even so, their "reality" was assured in advance by their emergence from statistical analysis. The scientific challenge lay in establishing their correct identities: 'technically speaking, they stand in need of reification' (Ferguson and Maclean, op. cit.:122). Accordingly, the authors conjecture that the first principal component represents 'general language difficulty', but remain doubtful about the second ('something to do with contextual support and rhetorical organization' (ibid.)).

Here again, it is important to be clear what the issue is. It is not in question that for the practical purposes (i.e. efficiency of text grading) that are the experimenters' immediate object (ibid.:118), discarding the unreliable and hardly quantifiable variables that contributed least to the overall assessment of difficulty is obviously sensible. Nor are the statistical procedures necessarily suspect in themselves. The problem appears when these procedures are used in the analysis of the underlying causes of the difficulty judgements, simultaneously to provide a conceptual model of the judgements (they "really" consist of two components), and empirical evidence to support it (their occurrence in this study), for in this way the method tends simply to confirm the validity of its own artefacts. Moreover, instead of achieving clarity of understanding, we are left facing a paradox: the statistical method is taken to have delivered a deeper,

more real picture of the relevant cognitive activity than the consciously evolved, subtle descriptions of the agents themselves, yet it proves, on inspection, to be devoid of interesting content. To the writers, the natural solution is to indicate the need to refine and elaborate their statistical analyses (ibid.:123). Without a theoretical model that will permit interpretation of the results independently of the method used to derive them, however, these refinements will be of little use.

1.4 Conclusion

The studies examined here illustrate in different ways the readiness of experimenters to use inferential statistical methods, independently of a theoretically conceived research design, to do the work of conceptual analysis, by-passing experienced judgement, and licensing the affirmation of general theoretical conclusions. The remainder of this paper considers these misconceptions in greater detail. The following section first outlines doubts raised about the place of significance testing in research in the behavioural and social sciences.

2. Questioning the significance test

2.1 The emergence of doubts

Insisting that results achieve a specified level statistical significance (e.g .01) has sometimes been used by editors as a means of preventing the literature from overflowing with spurious studies (see, for example, Melton quoted in Bakan 1966:426f). Yet by itself this does not hold back the tide. If anything, it inclines experimenters to publish claims for hunches apparently "confirmed", but to discount null hypotheses left unrejected; while if subsequent evidence then points to the truth of H_0, this will tend to go unreported (Carver 1978:396). The result, as noted above, is that significant progress is publicly announced where the physical sciences might at best see only 'private clues for future exploration' (Hogben 1970:19). More generally, attaching undue emphasis to statistical significance encourages experiment where none is justified, just because the test is easy to apply and creates an impression of scientific objectivity. So, for example, Venezky's survey of research into reading instruction published in the United States notes that the bulk of it is composed of meaningless statistical exercises, 'fishing expeditions ... almost random searches for relationships, unanchored by any theoretical frameworks and often unbothered by the limitations of the methods employed' (Venezky 1984:17); 'an enduring testimony to the patience of the American printer and the vulnerability of American forests' (ibid.).

Similar misgivings surfaced during the 1960's in the American psychological research literature, prompting a re-examination of the role played in it by statistical significance testing, and leading to conclusions such as Lykken's that

> statistical significance is perhaps the least important attribute of a good experiment; it is *never* a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence - or that an experimental report ought to be published.
>
> (Lykken 1968:158; original emphasis)

The particular object of criticism was professional ignorance of the logic and interpretation of the tests, and the widespread readiness to make invalid inferences based on them (Rozeboom 1960; Bakan 1966; Meehl 1967; Lykken 1968), a readiness for which empirical studies have provided evidence (see, for example, Rosenthal and Gaito 1963; Kahneman et al. 1982). The discussion extended to philosophical misgivings about the status of statistical inference in a scientific paradigm, reflecting disputes, described by Hogben (op. cit.), of longer standing within statistical theory itself about the true purpose and scope of the test. The psychological papers cited above, together with others from sociology and elsewhere, have been collected by Morrison and Henkel, who summarize the general consensus:

> The significance test as typically employed in behavioural science is bad statistical inference, and ... even good statistical inference in basic research is typically only a convenient way of sidestepping rather than solving the problem of scientific inference.
>
> (Morrison and Henkel 1970:xi)

The same conviction regarding educational research has since been forcefully expressed by Carver:

> The emphasis on statistical significance over scientific significance in educational research represents a corrupt form of the scientific method. Educational research would be better off if it stopped testing its results for statistical significance.
>
> (Carver 1978:378)

There is little sign, it must be said, that research practices in these fields have responded to such criticism; obstacles to adopting the solution Carver proposes are discussed in section 4.2.

2.2 The case of applied linguistics

Examining the issues raised by these critics as they relate to research in applied linguistics will highlight aspects of the relationship between probabilistic techniques and other kinds of inductive reasoning that are easily glossed over, and dispel any expectation that an empirical approach will in itself lead to clearer understanding, or that 'if our treatment of our subject matter is mathematical it is therefore precise and valid' (Bakan op. cit: 437). Far from being of exclusively philosophical interest, this activity should be seen, in Hogben's words, as 'the birthright and duty of every scientific worker who subjects his data to ... statistical inference' (Hogben, op. cit.:14-15).

Nevertheless, few applied linguists, at least among those from "arts" backgrounds, will feel qualified to evaluate the statistical techniques taught as the routine practice of the discipline, still less their adequacy within the experimental paradigm. It is easier to trust expert assurances that they "work", as one can learn to drive without understanding about cars. Since, moreover, the results turned out by these techniques will take the form preferred by the research community, it may never seem urgent to criticize their presuppositions. By these means, researchers are socialized into treating the statistical significance test as a paradigm of scientific rationality and ready-made inferential device sufficient to all normal purposes. Indeed, doubts of the kind expressed in

psychology, in education, and in sociology have hardly touched applied linguistics, which has by and large persevered in regarding statistical significance as a distinguishing mark of good experiments, and as a means of assessing the theoretical value of results. Some of the more detailed implications of this are pursued below.

There are, of course, reasons other than ignorance or force of habit to account for it. Researchers whose interests lie in aspects of learner or teacher behaviour need practical instruments in order to get on with their jobs. To the extent that they are concerned with theory, it is less likely to be with the aim of testing it, than of using it to solve a problem or illuminate a phenomenon. This may reasonably be thought of as the everyday life of any research field (see Putnam 1981a:70ff). Secondly, as Rozeboom pointed out in the case of psychology, rational argument may have made little impression on the research community just because, consciously or not, it has 'never taken the [null hypothesis test] method seriously anyway' (Rozeboom 1960:424). 'Who has ever given up a hypothesis just because one experiment yielded a test statistic in the rejection region?' (ibid.). Conversely, there is likely to be little pressure to criticize closely a method which helps to legitimate a favoured (and funded) line of enquiry.

3. Misinterpretations of p

3.1 Inference from experiment

We might expect that the logic of Popper's falsification criterion (see, for example, Popper 1959) would by now have displaced more or less naive claims, such as the one examined in 1.1, that one-off experiments had "confirmed", or "lent support to" some piece of theory, or "proved" the success of some treatment. According to Popper, confirmation by itself can add nothing to a theory. In his view, it is falsifiability (or refutability) that provides the criterion for demarcating empirical (i.e. scientific) theory from non-empirical theory. Thus, in general, for a hypothesis H_1, deduced from theory T, of the form "T implies q" (where q is some kind of observable phenomenon), attempting to observe q (the confirmatory approach) will allow no conclusion about the truth or falsity of T. Or, as Lakatos expresses it, 'it is no success for Newtonian theory that stones, when dropped, fall towards the earth, no matter how often this is repeated' (Lakatos 1978: 6). On the other hand, establishing that q is not the case logically entails that T is false. Thus, Popper argues, it is by refutation, not confirmation, that science ideally seeks to increase the store of well-grounded theoretical knowledge.

If this is so, the rarity of deliberate attempts at falsification in applied linguistics (and elsewhere) needs to be explained. It may not simply be that applied linguists are more eager to pass off speculation as fact than the honest Popperians in other disciplines. There are at least two more compelling possibilities. First, as Rozeboom's remark (quoted in 2.2) suggests, in the day to day conduct of science there is nothing so clear-cut about falsification as logic alone should require. Cumulative research would be an impossibility if a single negative result could overthrow a theory. For one thing, experimenters in every discipline generally hope to prove their hypotheses, not refute them, and quite naturally try to preserve them in the face of "recalcitrant instances". Refutation alone will tell us nothing about what is true. For another, for the purposes of normal science, theories (regarded as laws) will be able to draw on whole networks of protective "auxiliary hypotheses", statements of contingencies both predictable and

unknown, to which adjustments may be made without disturbing their core assumptions (cf. Lakatos, op. cit.:49ff; Putnam op. cit.).

The second, in the present context more pertinent, explanation is that the statistical significance test procedure actually promotes the quest for confirmation, by attaching no clear role to refutation; discarding an H_1 because it fails to reach (say) $t_{crit}=0.05$ is by no means simply refutation in statistical guise. As suggested in section 1.1, by appearing to pass the burden of responsibility for determining the truth of a theory away from the experimenter to the inferential structure of the test itself, the significance test procedure may even seem to hold out a valid means of verification, a tempting short-cut to what Bakan calls the experimenter's 'dream, fantasy or ideal', namely automatic inference (Bakan, op. cit.:430). The logic of hypothesis testing, reduced to the exercise of a simple binary decision, makes it easy to conflate a decision to reject H_0/accept H_1 with a decision to "accept" the truth of T itself.

Suppose, however, T states that invisible spirits determine whether coins fall heads or tails when tossed (cf. Bakan, op. cit.:425) and an experiment is then conducted in which, following appropriate invocation of the "heads" spirit, an unbiased coin falls heads ten times in a row. The probability (p) of such an outcome is $(½)^{10}$ or less than 0.001. Remarkable though this may be, it will be unlikely to strengthen most normal observers' confidence with respect to T, a confidence which would remain unchanged no matter how many times a "confirming instance" was achieved: ten throws would lend T neither more nor less support than a hundred - both results would lend it no support at all. While T's implausibility makes this obvious, the fact is a matter of Popperian logic, showing merely that 'it is easy to obtain confirmations or verifications for nearly every theory - if we look for confirmation' (Popper, cited in Wason 1977:307). Yet in terms of the statistical procedures we regularly depend on, $p<0.001$ is "highly significant". Adopting .05 as an acceptable measure of risk that the result might turn up by chance under the null hypothesis (H_0) that tails and heads occur with equal frequency, it will be clear that even five throws ($p=.031$) achieves "significance", entailing rejection of H_0. Manifestly, the decision to reject H_0 can carry no implication about the truth of T.

If this fact is clear to all users of significance tests, they need never risk overstating the importance of experimental results. The coin experiment would certainly not persuade us that it is 99.9% probable that spirits do in reality determine the toss of a coin, or again that a probability of only .1% attaches to H_0. The tendency to think either of these things is a manifestation of what Carver calls the 'valid research hypothesis fantasy' (Carver op. cit.:386). The power of this fantasy becomes easier to understand when a theory happens to be intrinsically more plausible, as theories in applied linguistics often are. A result such as that illustrated in section 1.1 offers quite strong psychological inducement to read (1-p) - or, in this case, the published T-ratios (Hafiz and Tudor op. cit.:8, table 2) - as a "confirmation-index" for the writers' theoretical assumptions. As Meehl points out (1967:107), professional test users are unlikely to fall into such an elementary error when it is stated baldly; nevertheless, the tendency to quote very low p values as if by their lowness they contributed substantively to the corroboration of T is widespread, and reinforced by the habitual use of the terms "significant" and "highly significant" to characterize them (cf. Carver 1978:386). Hatch and Farhady fail to warn us clearly away from this interpretation, even while apparently cautioning against over-confidence: 'If we claim significance at the .001 level (and in social sciences that's almost bragging about how sure we are!), there is still 1 chance in

1,000 we might be wrong ...' (Hatch and Farhady 1982:106); by 'wrong' they presumably mean 'committing a Type I error' (see below); but it is hard to resist the implication that there are 999 chances in 1,000 that we (and our theory) are right.

Furthermore, it will be observed that rejecting the coin result as valid evidence for T on the grounds that it obviously must have arisen by chance (etc.) is equivalent to accepting the result of the reading programme above as evidence for the Input Hypothesis, on the grounds that the idea seems plausible. Both cases involve stepping outside the test procedure to draw on reasons derived from other sources, such as our knowledge of the world, our expert judgement, or our commitment to a theoretical position: to the extent that such sources are available and decisive (as in science they should be), we do not need statistical significance tests as "decision mechanisms" at all.

3.2 Chance

p expresses the probability of our committing a Type I error: i.e. of falsely rejecting H_0. There is nevertheless a natural inclination among experimenters, reinforced, as Carver shows, by the writers of introductory statistical texts, to regard p as a statement of the 'odds against chance' (cf. Carver, op. cit.: 383). In its eagerness to get from statistical significance to corroborating conclusion, this is cognate with the belief in automatic inference. If p is held to express how likely it is that the result may have turned up by chance, its reduction to negligible levels implies that something substantial has been caught (presumably evidence for T) in the experimental net. In reality, of course, the significance test is premised on the truth of H_0, i.e. that chance ("sampling error") alone accounts for the observed result (and, it is worth noting, this assumption must be adhered to in practice by sampling at random from the population in question; if this is not done, for example where "convenience" samples are used, like the local school classes in the Hafiz and Tudor study, statistical significance can have no serious meaning (cf. Hewitt 1982:16)).

However, it is one thing to accept that results which are improbable under H_0 will occasionally turn up, so that over time a small percentage of published experiments will contain Type I errors, i.e. wrongly accept H_1 (ideally 5%, with $p=.05$ as the acceptance criterion, although pressure to select significant and ignore non-significant results will tend to push the number up). It is another, and something we cannot usually hope to determine, to say what the odds are in any given instance that a Type I error has occurred. In other words, the logic of significance testing allows us only to talk about central tendencies in the population of experiments, not about single instances. For this reason alone (independent of logical considerations) it would always be wrong to interpret a single result in the acceptance region as support, etc. for a hypothesis. Moreover, without considering the power of experiments, it is impossible to guess the likelihood of their having uncovered a "genuine" phenomenon, however minute the level of significance achieved (see 3.4 below).

3.3 Replication

One hallmark of a strong scientific theory is its resilience under repeated experiment. It must hold universally, subject to the calculable influence of other variables (plus various simplifying assumptions), not just for some favoured group of experimenters, or in some privileged location. If applied linguistics claims to deal with universal principles of language learning in this sense, its theories, too, should withstand

replication. In fact, the necessity is the more urgent, in one sense, given the relatively high proportion of spurious results the significance test procedure will instruct us to accept (presumably the spirit theory for coins will not survive a second trial).

In practice, replication tends to be neglected. It is true there is a genuine difficulty, faced by any behavioural discipline, of knowing exactly how replication is to be understood; for although physical laws may be independent of time, place, culture, etc., human behaviour clearly is not, so that it will be unreasonable to hope for more than a rough identity of conditions between the different occasions on which the "same" experiment is performed. But, as with refutation, the main reason appears to be that significance testing assigns no role to it; in fact, the test rules out replication a priori. As Bakan points out, the 'once-ness' of an experiment is a condition of the inferential model on which the test as Fisher conceived it is based; its logic presupposes an infinite hypothetical universe, of which the actual experiment represents a random sample, and it will be undermined by replication unless the probabilities are adjusted so as to treat both as a single entity (Bakan, op. cit. 424-5; cf. Hagood 1970:67). Thus, a succession of experiments designed to test the same theory, each achieving statistically significant results, cannot be regarded as a substitute for replication, whatever the temptation to do so. Conversely, we may take the view that a given set of behavioural data cannot be separated into its universal essence, and the effect of other variables and simplifying assumptions just referred to, in other words that it is uniquely shaped by its context and therefore non-replicable. In this case, if we test for significance, we shall need a clear understanding of just what 'infinite hypothetical universe' is intended (Hagood op. cit.:70).

3.4 Power

The difficulty of knowing what, if anything, is a substantive phenomenon in our experiments, and of being sure that it is such a phenomenon that an experiment has uncovered, raises the further difficult question of experimental power. In the natural sciences power will be a matter of instrumentation: finer scales, better lenses, etc. In the human sciences, including applied linguistics, it will depend principally on strength of experimental design and on sample size.

For practical purposes, if we want to discover the existence of real entities, rather than the non-existence of unreal ones, experimental power should interest us. Yet the significance procedure disregards it, except insofar as the prior determination of a critical level of significance is a trade-off between the acceptability of Type I errors and those of Type II (i.e. failure to reject a false H_0). While we can try to decide for ourselves whether an experimental design is valid, the fact remains that emphasis in published results on exceptionally small values of p diverts attention away from weaknesses. Carver calls this the 'replicability or reliability fantasy' (Carver: op. cit.:385). As long as (1-p) is taken, consciously or otherwise, to express the reliability of the result obtained, it will appear, quite unjustifiably, to validate *post hoc* whatever design has been used: crudely, that if it has a "highly significant" label attached to it, it must have been a good experiment. This also helps to reinforce the idea that statistical significance can stand in lieu of replication, with progressive research in a given field seen as the accumulation of results so labelled (cf. above).

Why this will not do will be discussed further in a moment. But the reason for it no doubt reflects the relative ease of calculating statistical significance, as against the

relative difficulty of calculating experimental power. In applied linguistics experiments, "noise" levels are generally high, so that there is no sure way of knowing if the effect observed is "really" the effect being looked for. Unlike typical scientific variables, which may be predicted, isolated and measured with great accuracy, the variables which interest us (L2 proficiency, reading comprehension, difficulty judgements, etc.) depend critically on the validity of a series of secondary inferences. They may be spoken of as independent entities - like 'sheer comprehension', for example, which the experimenter hopes to distinguish from memory, inference, deduction, reasoning, intelligence, etc. (Carroll 1972:3ff) - but their identities have themselves emerged from statistical procedures with quantities (such as test scores) whose interpretation is open to debate, and interact non-randomly with a host of variables in the educational, psychological and cultural backgrounds of the subjects. Instead of making point predictions, our hypotheses deal with directional tendencies in population means, inferable only on the basis of (frequently small) samples; there is no ready way of predicting how big observed effects should be, and no independently interpretable scale on which to represent them.

It would be wrong to expect obtained levels of significance to fill all, or any, of these requirements, given the ease with which they can be manipulated by the experimenter, especially where sample size is concerned (see below). On the other hand, for normal, "messy" experimental situations power cannot be calculated in advance. Therefore, we cannot know how often we fail to detect a difference that really exists, in other words commit a Type II error. The proportion of Type II errors in the population of applied linguistics experiments must nevertheless be presumed to be rather high, since the probability of Type II error - which we may call p_2 - will be inversely proportional to the probability of Type I error, p_1 (or, simply, p): i.e. the fewer Type I errors we allow (by being reluctant to reject H_0), the more Type II errors we will make (by letting through H_0s that are in fact false), and vice versa. Our aim will be to tighten experimental design towards the ideal point of no noise, where $p_2=0$, but, as we have seen, noise is an irreducible property of the variables we want to investigate. Power, expressed by $(1-p_2)$, must therefore frequently be low.

According to Tversky and Kahneman (1982) in a study of how probabilistic data are interpreted by those who use them, the worst of the 'self-defeating' and 'pernicious' consequences of low experimental power is not so much the valid but discarded hypotheses strewn along the pathway of research, as the readiness on the part of experimenters, for which they quote evidence from a survey of research psychologists, to explain noise (ibid.:27), to seek causal explanations for unexpected differences between the results of an experiment and its attempted replication, when it is quite beyond the power of the experiment to resolve them into true and chance effects. In this way, experiments constantly add spurious facts to the repository of knowledge, and equally spurious theories may be evolved to explain them.

3.5 Power and "corroboration": a paradox

A paradox discussed by Meehl (1967) concerning the nature of experimental power illustrates the true distance between hypothesis testing in behavioural and natural sciences. It rests on the reasonable assumptions that (1) the aim of any experimenter will be to improve experimental power towards 100%; and that (2) H_0 is almost always false in any population; in other words that any treatment (e.g. extensive reading) or criterion of classification (e.g. sex, father's religion, etc.) will have <u>some</u> influence on

output measures that will be detectable given sufficient power, or, which is the same thing, a large enough sample (see also 3.6). Meehl's argument adopts, as a limiting case, the further assumption (3) that all our theories are equally unlikely. If assumption (2) is true, it follows, quite independently of other considerations, that, over the infinite set of possible experimental or quasi-experimental situations in (say) applied linguistics, there will be a non-zero difference between "experimental" and "control" groups on the variable of interest in practically every case (and for the purpose of this argument it does not matter which group is so designated). Assuming that these differences are normally distributed, the difference observed will be in the direction that favours the experimental group in 50% of cases. Let this outcome (again arbitrarily) be called "success", and that in the other 50% of cases, in which the difference favours the control group, be called "failure". If all these experimental situations are then paired off randomly with theories drawn from the infinite set of real or potential theories in applied linguistics, there is a 50% probability that a theory, however wrong "in the state of nature", will find itself paired with a "successful" experiment. In other words, the consequence of our experimental method is to yield a prior probability equal to 50% of finding experimental "support" for any of our theories (ibid.:113).

Meehl is at pains to stress that this is a limiting case, 'a lower bound on the success-frequency of experimental "tests"'(ibid.:111), given assumption (3) above, and assuming "perfect power" (i.e. certain detection of any difference that exists). He therefore concludes that, paradoxical as it may seem, any attempt to increase the power of experiments in the real world will only make the "observational hurdle" for a theory easier to overcome, by bringing the probability of achieving "corroboration" ever closer to 50%, even where the theory in question is intrinsically worthless. In the natural sciences, by contrast, increasing experimental power achieves just the reverse: as calibrations and measurements gain in precision, so theories are forced to pass progressively more stringent tests, reducing towards zero the chances of survival for any but the very fittest.

When combined with the ever-present temptation, discussed earlier, to confuse rejection of H_0 with confirmation of T, Meehl's paradox shows how easily even an apparently fruitful research programme might come to be based entirely on a self-perpetuating chain of flawed statistical inferences.

3.6 The fiction of the null hypothesis

The only conclusion that can legitimately be drawn from a statistically significant result is that there is a probability equal to the obtained value of p that H_0 was wrongly rejected. As Hogben comments:

> For what reason ... should [the researcher] be eager to take advantage of a test which can merely assign a low probability to erroneously asserting that the treatment is useless ... ? ... The terminal statement which the test procedure ostensibly endorses provides an answer (if any) devoid of operational value in the context of an experiment rightly undertaken to confirm a positive assertion suggested by prior information. Since the test procedure merely endorses the negation of a null hypothesis conceived within the straitjacket of the single infinite hypothetical population, the outcome will thus be an irrelevant decision or no decision at all.
>
> (Hogben, op. cit.:35)

14

It has been the object of the argument to this point to illustrate the hankerings that are widely felt for some surer mechanism to generate true substantive statements; hankerings which are readily but illicitly projected on to the significance test. No doubt these dangers are well-known: it would be an elementary mistake to attach to p any interpretation other than the one given above. It is unlikely that researchers will cling to just one of the false notions discussed; but they may be inclined, if only under the influence of common usage, at different times to fall into any of them. The most serious problem arises when the level of significance expressed by p is made to bear the weight of a decision to regard a piece of research as deserving further attention.

The thrust of Hogben's criticism here, however, is aimed less at the feebleness of assertion made possible by the significance test than at the fiction of the null hypothesis (Carver's 'straw man' (op. cit.:381)) upon which it is premised. Not only can we never claim to have "confirmed" H_0 on logical grounds, but to seek such confirmation would be irrational a priori. Researchers do not set up experiments in the belief that their treatment has no effect whatever, that the mean difference between an infinite number of sample scores drawn at random from an infinite population, treated and untreated, will be precisely 0.00. Given the complexity of the variables of interest to us, the probability of such a result is vanishingly small. It is safe to say that H_0 is never true. Indeed, for cases in which H_0 has not been rejected, it is common simply to repeat the experiment with a larger sample. Bakan refers to his own tests on data collected from 60,000 subjects: if N is large enough, almost any difference (e.g. east *vs* west of the Mississippi, north *vs* south, etc.) can be shown to be "significant" (Bakan, op. cit. 425; cf. also Meehl, op. cit.:109). Therefore the notion of random sampling under H_0, which is essential to the calculation of p, corresponds to no conceivable state of the experimental situation, either on a single occasion, or (much less) over time (cf. Hagood, cited in Hogben, op. cit.:40). In these circumstances it would be hard to make a convincing case on scientific grounds for persisting in its use.

4. Conclusion

4.1 Decision versus interpretation

These remarks may be enough to establish that significance testing is not in any sense just a version of scientific procedure tailored for a behavioural discipline. Even when it is approached as a probabilistic mechanism of strictly limited utility, however, there remain unresolved anomalies in its use, traceable, as Hogben shows, to profound disagreement at the level of statistical theory. If it is regarded as a decision test, triggering acceptance of hypotheses that achieve a pre-set level of significance, then its function should only be to ensure stability in the incidence of Type I error across the aggregate of experimental results. As such, it will admit no interpretation of results considered singly, and no attempt to equate levels of significance with an experimenter's strength of conviction. Yet there is a widespread tendency, already noted, for significance tests to be applied for the purpose of establishing degrees of belief with respect to single results. Worse, the two approaches are regularly combined without thought for their divergent implications, making it appear that a series of positive test decisions actually entails increasing conviction, even though such factors as the critical level of significance and the size of sample chosen are arbitrary or matters of convention.

This is no way for science to proceed. As Rozeboom argues:

> A hypothesis is not something like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true. ... While the scientist - i.e. the person - must indeed make decisions, his science is a systematized body of (probable) knowledge, not an accumulation of decisions.
>
> (Rozeboom, op. cit.:423; original emphasis)

Since experimenters have a duty to interpret their findings, and since their aim must be to establish a sound basis for the growth of knowledge in their field, the moral to be drawn is that significance tests, in any of their guises, are at best weak, at worst inappropriate and misleading. Lakatos, viewing these matters from the perspective of natural science, puts the point less charitably:

> One wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phoney corroborations and thereby a semblance of "scientific progress" where, in fact, there is nothing but an increase in pseudo-intellectual garbage.
>
> (Lakatos 1978:88, n.4)

4.2 Further implications

What would abandoning the test for statistical significance, or at least relegating it to a minor role in the analysis of data, mean for research in applied linguistics? The chief obstacle to dispensing with the significance test is that without it the research enterprise would seem to founder; not because there are no alternatives, but because that enterprise is, to a great extent, premised on the test and the kind of the knowledge it makes available. As Bakan has put it:

> [The test of significance] is profoundly interwoven with other strands of the ... research enterprise in such a way that it constitutes a critical part of the cultural-scientific tapestry. To pull out the strand of the test of significance would seem to make the whole tapestry fall apart.
>
> (Bakan 1966:428)

The failure of qualitative methods to make headway against inferential statistics may be attributable in part to the fact that the former are likely to be viewed as anecdotal, short of "scientific" rigour and inadequately generalizable; in short, as not conforming to the notion of properly constituted knowledge in the field. For this reason, there exists no established research discourse in applied linguistics to which such methods can contribute. But at the same time, no doubt, this state of affairs has been maintained by the prestige of the physical paradigm, by the tendency of our culture 'to view the exact sciences as the long-sought description of the "true and ultimate furniture of the universe"' (Putnam 1981b:15). It is against this background that the present discussion, echoing the work cited in section 2.1, has sought to put in doubt the assumed scientific rigour of significance testing, by showing that in many cases it is illusory and its implications readily misinterpreted. It has argued that the use not only of the logic but also of the language of "corroboration" derived from physical science creates the

impression of an empirical research programme progressing towards an ever clearer and better supported theoretical understanding, where no such impression may be justified.

It remains, however, that the special status of physical science naturally favours the belief that its representations are more nearly "true" in some absolute sense, so that it may still be taken for granted that better theoretical description of phenomena must mean, ultimately, closer approximation to the "objective" picture of the world delivered by physics. It is perhaps because the notion of "correspondence to the facts" is taken to be unproblematic that researchers even in the non-physical sciences have been able to develop highly sophisticated methodologies without giving equivalent attention to conceptual issues, treating methods as different kinds of tools, and choice among them as independent of the conceptualization of the empirical "facts" to be discovered. The further purpose of this argument has therefore been to suggest, on the contrary, that '"objects" do not exist independently of conceptual schemes. We cut up the world into objects when we introduce one or another scheme of description' (Putnam, op. cit.:52); that, as Hacking observes, 'a style of reasoning may determine the very nature of the knowledge that it produces' (Hacking 1981:143; cf. also his 1982:49ff).

For just this reason, it would be wrong to imply that probabilistic methods are intrinsically less valid than others. It is a matter of history that statistical modes of thought have increasingly been perceived as explanatory in the human sciences, and have made it possible to trace interesting relationships among phenomena. The very idea of "the human sciences" owes its possibility to advances in probabilistic techniques and the emergence of styles of reasoning associated with them in the nineteenth century (see, for example, Porter 1986, Stigler 1986, Hacking 1990). But to the extent that statistical methods are supposed, in the paradigm of physical science, to reveal (for example) the hidden facts of human cognitive operation, it is necessary to question the more or less automatic use that is made of them in our field.

Abandoning the significance test is not therefore just a methodological problem, or a matter of personal preference, as if we might replace it with something perhaps 'softer' and more congenial but in other respects continue with the work we are doing. If we accept that a methodology is (or necessarily implies) a style of reasoning, the change will essentially redefine that work itself, that is, the objects of research and the ways in which we think about them.

Acknowledgements

I should like to thank Robert Hill for his encouragement in the writing of an earlier version of this paper.

Notes

1. For a more circumspect analysis of the same results, see also Tudor and Hafiz (1989).

2. Perhaps it is unfair to discuss in detail research that is only reported in a working paper; however, attention here is directed less to its specific results than to the way the writers conceptualize one part of their project.

References

Bakan D. 1966. 'The test of significance in psychological research'. Psychological Bulletin 66/6: 423-37.

Carver R. 1978. 'The case against statistical significance testing'. Harvard Educational Review 48/3: 378-99.

Carroll J. 1972. 'Defining language comprehension: some speculations' in J. Carroll and R. Freedle (eds.) 1972. Language Comprehension and the Acquisition of Knowledge. New York: Wiley and Sons. 1-29.

Gould S. 1981. The Mismeasure of Man. New York: Norton.

Gregg K. 1984. 'Krashen's monitor and Occam's razor'. Applied Linguistics 5/2: 79-100.

Ferguson G. and J. Maclean. 1991. 'Assessing the readability of medical journal articles: an analysis of teacher judgements'. Edinburgh Working Papers in Applied Linguistics 2: 112-124.

Hacking I. 1981. 'Lakatos's philosophy of science' in I. Hacking (ed.) 1981. 128-143.

Hacking I. (ed.) 1981. Scientific Revolutions. Oxford: Oxford University Press.

Hacking I. 1982. 'Language, truth and reason' in M. Hollis and S. Lukes (eds.) 1982. Rationality and Relativism. Oxford: Blackwell. 48-66.

Hacking I. 1990. The Taming of Chance. Cambridge: Cambridge University Press.

Hafiz F. and I. Tudor 1989. 'Extensive reading and the development of language skills'. ELTJ 43/1: 4-11.

Hagood M. 1970. 'The notion of a hypothetical universe' in D. Morrison and R. Henkel (eds.) 1970. 65-78.

Hatch E. and H. Farhady 1982. Research Design and Statistics for Applied Linguistics. Cambridge, Mass.: Newbury House.

Hewitt G. 1982. 'A critique of research methods in the study of comprehension'. British Educational Research Journal 8/1: 9-21.

Hogben L. 1970. 'The contemporary crisis or the uncertainties of uncertain inference' and 'Statistical prudence and statistical inference' in D. Morrison and R. Henkel (eds.) 1970. 8-40. (Chapters reprinted from L. Hogben 1957. Statistical Theory. New York: Norton.)

Johanningmeier E. 1980. 'American educational research: applications and misapplications of psychology to education' in J. Smith and D. Hamilton (eds.) 1980. The Meritocratic Intellect: studies in the history of educational research. Aberdeen: Aberdeen University Press. 41-57.

Kahneman D., P. Slovic and A. Tversky (eds.) 1982. Judgement Under Uncertainty: Heuristics and Biases. Cambridge: Cambridge University Press.

Lakatos I. 1978. 'Falsification and the methodology of scientific research programmes' in J. Worrall and G. Currie (eds.) 1978. The Methodology of Scientific Research Programmes: Philosophical Papers of Imre Lakatos 1. Cambridge: Cambridge University Press. 8-101.

Lykken D. 1968. 'Statistical significance in psychological research'. Psychological Bulletin 70/3: 151-59.

McLaughlin B. 1987. Theories of Second Language Learning. London: Arnold.

Meehl P. 1967. 'Theory-testing in psychology and physics: a methodological paradox'. Philosophy of Science 34: 103-15.

Morrison D. and R. Henkel (eds.) 1970. The Significance Test Controversy. London: Butterworths.

Popper K. 1959. The Logic of Scientific Discovery. London: Hutchinson.

Popper K. 1979. Objective Knowledge: an Evolutionary Approach (revised edition). Oxford: Clarendon Press.

Porter T. 1986. The Rise of Statistical Thinking 1820-1900. Princeton: Princeton University Press.

Putnam H. 1981a. 'The "corroboration" of theories'. in I. Hacking (ed.) 1981. 60-79.

Putnam H. 1981b. Reason, Truth and History. Cambridge: Cambridge University Press.

Rosenthal R. and J. Gaito. 1963. 'The interpretation of levels of significance by psychological researchers'. Journal of Psychology 55: 33-38.

Rozeboom W. 1960. 'The fallacy of the null-hypothesis significance test'. Psychological Bulletin 57/5: 416-28.

Stigler S. 1986. The History of Statistics: the Measurement of Uncertainty before 1900. Cambridge, Mass.: The Belknap Press of Harvard University Press.

Tudor I. and F. Hafiz 1989. 'Extensive reading as a means of input to L2 learning'. Journal of Research in Reading 12/2: 164-178.

Tversky A. and D. Kahneman. 1982. 'Belief in the law of small numbers'. in D. Kahneman, P. Slovic and A. Tversky (eds.) 1982: 23-31.

Venezky R. 1984. 'The history of reading research' in P. Pearson (ed.) 1984. Handbook of Reading Research. New York: Longman. 3-38.

Wason P. 1968. 'On the failure to eliminate hypotheses. a second look' in P. Johnson-Laird and P. Wason 1977. Thinking: Readings in Cognitive Science. Cambridge: Cambridge University Press. 307-314.

Editor's note

Ferguson and Maclean were invited to respond to the comments on their work in this paper. Maclean does not feel it necessary to do so, Ferguson has indicated that he may respond later.